DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF TORONTO

CSC318S

THE DESIGN OF INTERACTIVE COMPUTATIONAL MEDIA

Lecture 12 — 25 February 1998

INTERACTION THROUGH SPEECH AND SOUND

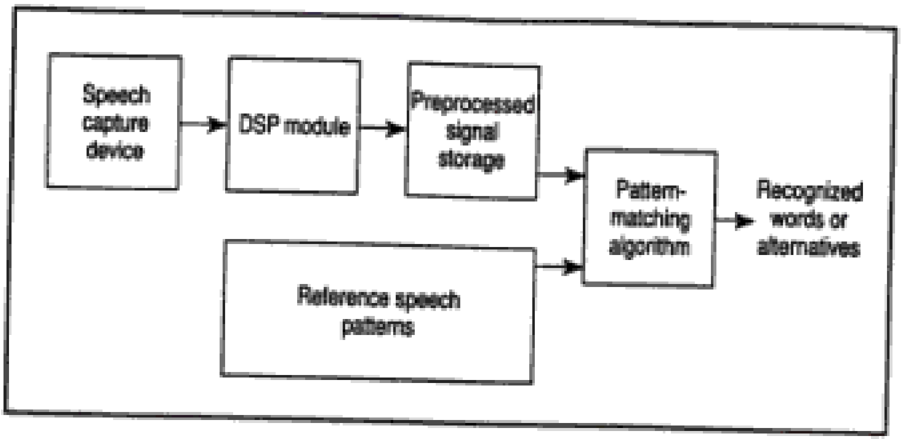
Ronald Baecker
Professor of Computer Science,
Electrical and Computer Engineering, and Management
University of Toronto

12.1 Input via speech recognition

Ideal applications
Hands busy or covered in "gunk"
Manual input already overloaded
Disabled users

Method of operation (Fig. 12.1)
Recognition vocabulary represented as stored patterns
Speech sampled and digitized
Waveforms or their parameters compared against patterns

Fig. 12.1 Components of a typical isolated word recognition system (BGBG, 1995, p. 548)


Speech capture device
DSP module
Preprocessed signal storage
Pattern-matching algorithm
Recognized words or alternatives
Reference speech patterns

Dimensions of success
Size of vocabulary: A few words to tens of thousands
Accuracy, recognition percentage: >>95%, >99%
Repeatability of performance
Cost
Speaker-dependent vs. speaker-independent
Training not required or easily trainable
Location of microphone
Acoustic environment, quiet or noisy environment
Discrete words or continuous speech

VIDEO — The OM System Spoken Language Interface
- *(Carnegie-Mellon University, SGVR 64, 1991)*
- Methods of user error correction
- Recognition architecture
- Usage of lexical and syntactic information (certain words & sentence structure are legal & therefore expected)

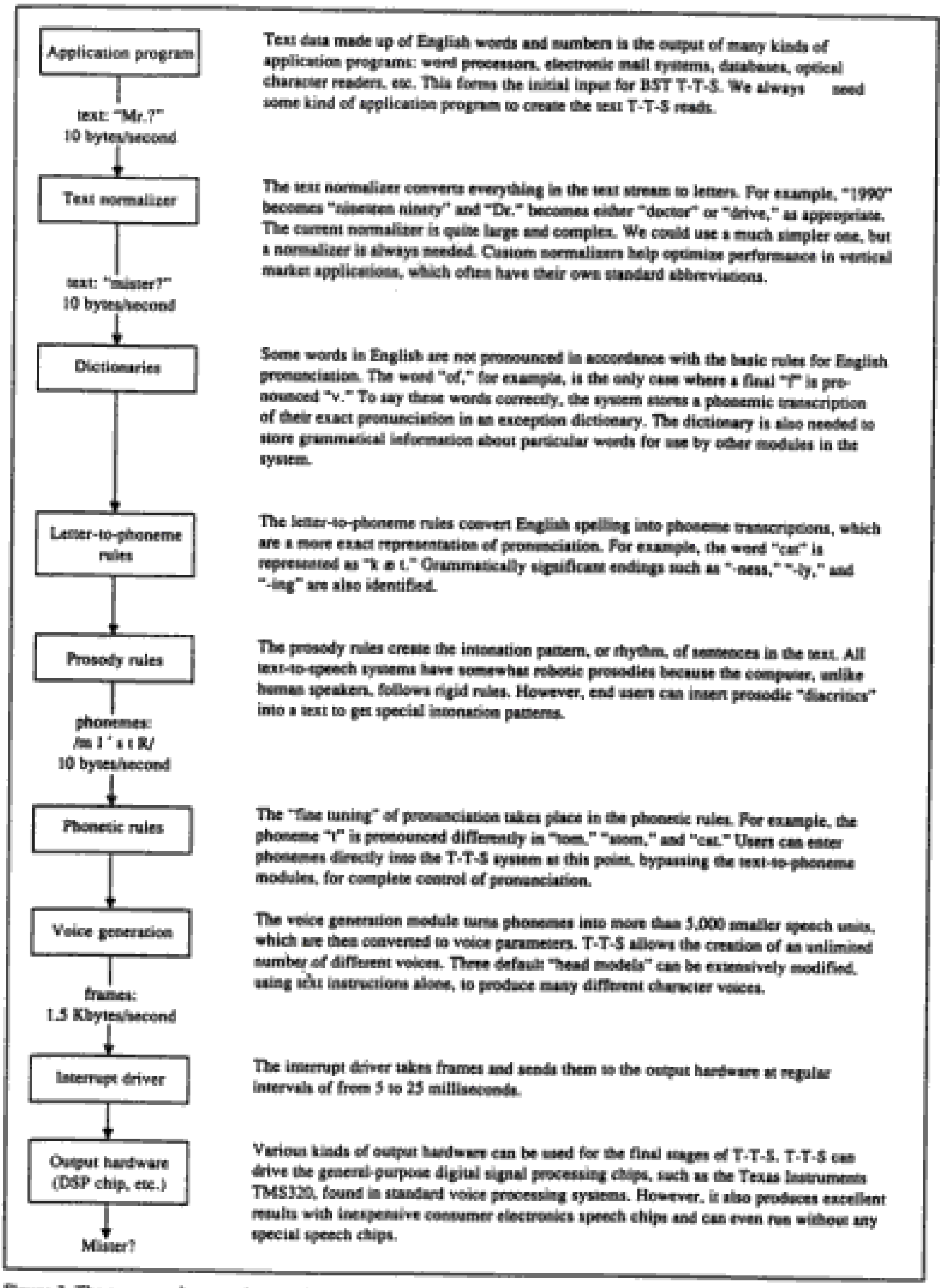
12.2 Output via speech synthesis

Why is the problem hard? Examples:
- How to pronounce “gh”?
    - No sound in “thorough”
    - “f” in “enough”
    - “g” in “ghost”
- How to pronounce “invalid”?
    - Not valid ==> Accent on second syllable
    - Disabled ==> Accent on first syllable
- How to stress (intonation)?
    - “I told you” means different things depending upon which word is stressed

Method of operation
- Digitized (stored) versus synthesized speech
- Synthesized speech
    - Phoneme-to-speech
    - Text (ASCII)-to-speech (Fig. 12.2)
- Retrieve or generate waveform, convert to analog, output

Dimensions of success
- Size of vocabulary
- Bandwidth, data rate
- Intelligibility
- Cost
- Naturalness
- Discrete words versus connected speech

Fig. 12.2 One method for conversion of text to speech (BGBG, 1995, p. 543)



Text data made up of English words and numbers is the output of many kinds of application programs: word processors, electronic mail systems, databases, optical character readers, etc. This forms the initial input for BST T-T-S. We always need some kind of application program to create the text T-T-S reads.

The text normalizer converts everything in the text stream to letters. For example, "1990" becomes "nineteen ninety" and "Dr." becomes either "doctor" or "drive," as appropriate. The current normalizer is quite large and complex. We could use a much simpler one, but a normalizer is always needed. Custom normalizers help optimize performance in vertical market applications, which often have their own standard abbreviations.

Some words in English are not pronounced in accordance with the basic rules for English pronunciation. The word "of," for example, is the only case where a final "f" is pronounced "v." To say these words correctly, the system stores a phonemic transcription of their exact pronunciation in an exception dictionary. The dictionary is also needed to store grammatical information about particular words for use by other modules in the system.

The letter-to-phoneme rules convert English spelling into phoneme transcriptions, which are a more exact representation of pronunciation. For example, the word "cat" is represented as "k æ t." Grammatically significant endings such as "-ness," "-ly," and "-ing" are also identified.

The prosody rules create the intonation pattern, or rhythm, of sentences in the text. All text-to-speech systems have somewhat robotic prosodies because the computer, unlike human speakers, follows rigid rules. However, end users can insert prosodic "diacritics" into a text to get special intonation patterns.

The "fine tuning" of pronunciation takes place in the phonetic rules. For example, the phoneme "t" is pronounced differently in "tom," "stom," and "cat." Users can enter phonemes directly into the T-T-S system at this point, bypassing the text-to-phoneme modules, for complete control of pronunciation.

The voice generation module turns phonemes into more than 5,000 smaller speech units, which are then converted to voice parameters. T-T-S allows the creation of an unlimited number of different voices. Three default "head models" can be extensively modified, using text instructions alone, to produce many different character voices.

The interrupt driver takes frames and sends them to the output hardware at regular intervals of from 5 to 25 milliseconds.

Various kinds of output hardware can be used for the final stages of T-T-S. T-T-S can drive the general-purpose digital signal processing chips, such as the Texas Instruments TMS320, found in standard voice processing systems. However, it also produces excellent results with inexpensive consumer electronics speech chips and can even run without any special speech chips.

Figure 3. The process of converting text into speech parameters in Berkeley Speech Technologies T-T-S system.

VIDEO — Talking to Machines (University of Wales, SGVR 88, 1993)

Speech input and output

Consequences of failure to anticipate user errors

Example of a real application

“Design principles”, but consider whether or not they are valid if:

- Users are very different, e.g., handicapped
- Machines and context of use is very different, e.g., portable PDAs rather than desktop machines

12.3 Recent advances in speech I/O

Word spotting for speech skimming

Speeding up digitized speech output

Multi-modal input and output

Use together with other techniques, such as voice output, language understanding, large screen display, gestural input

VIDEO — PUT THAT THERE (MIT, 1981, SGVR 13)

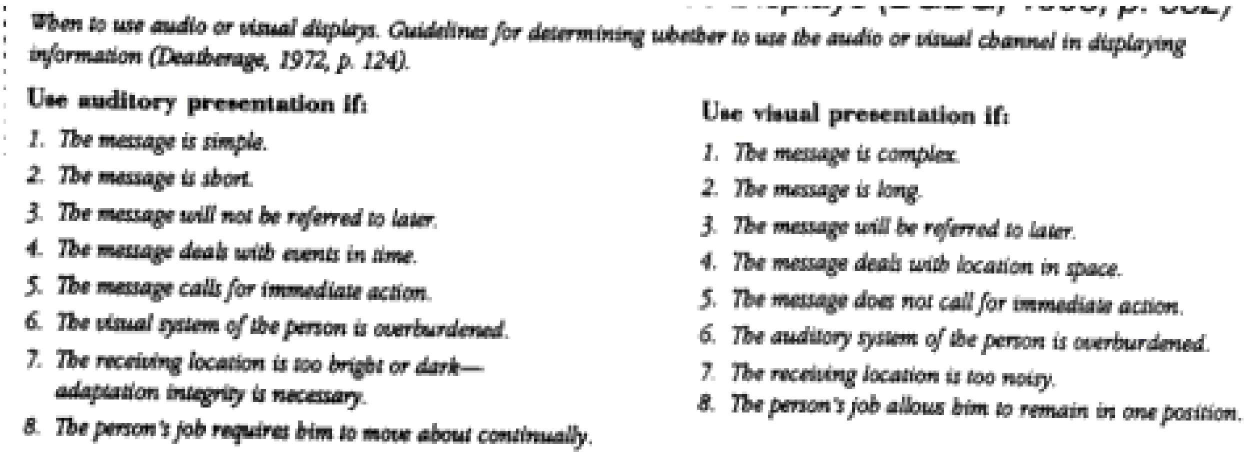
12.4 Auditory output

Roles for auditory displays

- Alarms and warnings
- Status and monitoring indicators
  - e.g., feedback from control inputs
- Messages and data (perhaps encoded)
  - e.g., responses to user queries

Visual versus auditory displays (Fig. 12.3)

Fig. 12.3 When to Use Audio or Video Displays (BGBG, 1995, p. 532)



When to use audio or visual displays. Guidelines for determining whether to use the audio or visual channel in displaying information (Deatherage, 1972, p. 124).

Use auditory presentation if:

1. *The message is simple.*
2. *The message is short.*
3. *The message will not be referred to later.*
4. *The message deals with events in time.*
5. *The message calls for immediate action.*
6. *The visual system of the person is overburdened.*
7. *The receiving location is too bright or dark—adaptation integrity is necessary.*
8. *The person's job requires him to move about continually.*

Use visual presentation if:

1. *The message is complex.*
2. *The message is long.*
3. *The message will be referred to later.*
4. *The message deals with location in space.*
5. *The message does not call for immediate action.*
6. *The auditory system of the person is overburdened.*
7. *The receiving location is too noisy.*
8. *The person's job allows him to remain in one position.*

12.5 Non-speech audio output

Motivation

Consider role of sound in video games, driving
Warnings (e.g., sound of blowout)
Status indicators (e.g., revving engine)
Feedback (e.g., grinding gears)

VIDEO — Sonic Finder (Gaver, UCSD and Apple, mid-80s)

Hear the trash can through a “tinny crash”
Hear amount of space on disk through reverberation
Hear status of scrolling through ascending or descending tones

Issues

Appropriate acoustic design
Storage requirements or real-time processing
Acoustic pollution

VIDEO — LogoMedia (DiGiano, Baecker, 1993)

Use of sound in software visualization